QUANTITATIVE COMPARISONS BETWEEN FINITARY POSTERIOR DISTRIBUTIONS AND BAYESIAN POSTERIOR DISTRIBUTIONS

FEDERICO BASSETTI

ABSTRACT. The main object of Bayesian statistical inference is the determination of posterior distributions. Sometimes these laws are given for quantities devoid of empirical value. This serious drawback vanishes when one confines oneself to considering a finite horizon framework. However, assuming infinite exchangeability gives rise to fairly tractable *a posteriori* quantities, which is very attractive in applications. Hence, with a view to a reconciliation between these two aspects of the Bayesian way of reasoning, in this paper we provide quantitative comparisons between posterior distributions of finitary parameters and posterior distributions of allied parameters appearing in usual statistical models.

1. INTRODUCTION

In the Bayesian reasoning the assumption of infinite exchangeability gives rise to fairly tractable *a posteriori* quantities, which is very attractive in real applications. If observations form an infinite exchangeable sequence of random variables, de Finetti's representation theorem states that they are conditionally independent and identically distributed, given some random parameter, and the distribution of this random parameter is the center of the current Bayesian statistical inference. The theoretical deficiency of this approach lies in interpreting these parameters. In fact, as pointed out for the first time by de Finetti (see [5],[6] and also [4]), parameters ought to be of such a nature that one should be able to acknowledge at least the theoretical possibility of experimentally verifying whether hypotheses on these parameters are true or false. A closer look to the usual Bayesian procedures shows that Bayesian statisticians often draw inferences (from observations) both to empirical (i.e. verifiable) and to non empirical hypotheses. To better understand this point, it is worth stating a more complete formulation of the already mentioned de Finetti's representation theorem: *A sequence* $(\xi_n)_{n\geq 1}$ *of random elements taking values in some suitable measurable space* $(X, \mathcal{X})$ *(e.g. a Polish space), is exchangeable if and only if the empirical distribution*

$$\tilde{e}_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}(\cdot)$$

converges in distribution to a random probability $\tilde{p}$ *with probability one and the* ξ_n*s turn out to be conditionally independent given* $\tilde{p}$*, with common distribution* $\tilde{p}$*.* Hence, it is $\tilde{p}$ that takes the traditional role of parameter in Bayesian modeling. However, since $\tilde{p}$ is a limiting entity of mathematical nature, hypotheses related to it might be devoid of empirical value. It

Key words and phrases. de Finetti's theorem, Dudley metric, empirical distribution, finitary Bayesian inference, finite exchangeability, Gini-Kantorovich-Wasserstein distance, predictive inference, quantitative comparison of posterior distributions.

AMS classification: 62C10,62F15,60G09

Research partially supported by Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR grant 2006/134526) the IMATI (CNR - Pavia, Italy).

is clear that this drawback vanishes when one confines oneself to considering a finite horizon framework, in which $\tilde{e}_N$, that is always (at least ideally) observable, takes the place of $\tilde{p}$. In this way one preserves the hypothesis of exchangeability, which is quite natural in many statistical problems, but one avoids the problem of assessing probability law to unobservable entities. In particular, in this context, the conditional distribution of the empirical measure $\tilde{e}_N$ given $\xi(n) := (\xi_1, \dots, \xi_n)$ $(n < N)$ takes the place of the conditional distribution of $\tilde{p}$ given given $\xi(n)$, i.e. the usual posterior distribution of the Bayesian (nonparametric) inference.

Even if, in view of de Finetti's representation, the parameter corresponding to the so–called "unknown distribution" (i.e. $\tilde{p}$) is the limit, as $N \to +\infty$, of empirical distribution, it should be emphasized that in the Bayesian practice two conflicting aspects sometimes occur. On the one hand, statistical inference ought to concern finitary and, therefore, observable entities whereas, on the other hand, simplifications of a technical nature can generally be obtained by dealing with (parameters defined as function of) the "unknown distribution" $\tilde{p}$. Hence, it is interesting to compare the conditional distribution of $\tilde{e}_N$ given $\xi(n)$ with the conditional distribution of $\tilde{p}$ given $\xi(n)$, when $(\xi_k)_{k\geq 1}$ is an infinite exchangeable sequence directed by $\tilde{p}$. This is the aim of this paper, that can be thought of as a continuation of the papers [2] and [3], where specific forms of (finitary) exchangeable laws have been defined and studied in terms of finitary statistical procedures.

The rest of the paper is organized as follows. Section 2 contains a brief overview of the finitary approach to statistical inference together with some examples. Sections 3 and 4 deal with the problem of quantifying the discrepancy between the conditional law of $\tilde{e}_N$ given $\xi(n)$ and the conditional law of $\tilde{p}$ given $\xi(n)$.

To conclude these introductory remarks it is worth mentioning [7], which, to some extent, is connected with our present work. In point of fact, in [7], Diaconis and Freedman provide an optimal bound for the total variation distance between the law of $(\xi_1, \dots, \xi_n)$ and the law of $(\zeta_1, \dots, \zeta_n)$, $(\xi_1, \dots, \xi_N)$ being a given finite exchangeable sequence and $(\zeta_k)_{k\geq 1}$ a suitable infinite exchangeable sequence.

2. Finitary statistical procedures

As said before, we assume that the process of observation can be represented as an infinite exchangeable sequence $(\xi_k)_{k\geq 1}$ of random elements defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in a complete separable metric space (X, d), endowed with its Borel σ–field $\mathcal{X}$. Let $\mathbb{P}_0$ be a subset of the set $\mathbb{P}$ of all probability measures on $(X, \mathcal{X})$ and let $t : \mathbb{P}_0 \to \Theta$ be a parameter of interest, Θ being a suitable parameter space endowed with a σ–field. From a finitary point of view, a statistician must focus his attention on empirical versions $t(\tilde{e}_N)$ of the more common parameter $t(\tilde{p})$.

It might be useful, at this stage, to recast the decision theoretic formulation of a statistical problem in finitary terms. Usually one assumes that the statistician has a set $\mathbb{D}$ of *decision rules* at his disposal and that these rules are defined, for any $n \leq N$, as functions from $\mathbb{X}^n$ to some set $\mathbb{A}$ of actions. Then one considers a *loss function* L, i.e. a positive real–valued function on $\Theta \times \mathbb{A}$, such that $L(\theta, a)$ represents the loss when the value of $t(\tilde{e}_N)$ is θ and the statistician chooses action a. It is supposed that

$$r_N(\delta(\xi(n))) := \mathbb{E}[L(t(\tilde{e}_N), \delta(\xi(n)))|\xi(n)]$$

is finite for any δ in $\mathbb{D}$ and, then, $r_N(\cdot)$ is said to be the *a posteriori Bayes risk* of $\delta(\xi(n))$. Moreover, a *Bayes rule* is defined to be any element δ_{FB} of $\mathbb{D}$ such that

$$r_N(\delta_{FB}(\xi(n))) = \min_{\delta\in\mathbb{D}} r_N(\delta(\xi(n)))$$

for any realization of $\xi(n)$. We shall call such a Bayes rule *finitary Bayes estimator* in order to distinguish it from the more common Bayes estimator obtained by minimizing

$$r(\delta(\xi(n))) := \mathbb{E}[L(t(\tilde{p}), \delta(\xi(n)))|\xi(n)].$$

While the law of the latter estimator is determined by the posterior distribution, that is the conditional distribution of $\tilde{p}$ given $\xi(n)$, the law of a finitary Bayes estimator is determined by the "finitary" posterior distribution, that is the conditional distribution of $t(\tilde{e}_N)$ given $\xi(n)$. A few simple examples will hopefully clarify the connection between the finitary Bayesian procedures and the usual Bayesian ones. In all the examples we shall present, observations are assumed to be real–valued, that is $(X, \mathcal{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the space of actions is some subset of $\mathbb{R}$ and the loss function is quadratic, i.e. $L(x, y) = |x - y|^2$. It is clear that, under these hypotheses,

$$\delta_{FB}(\xi(n)) = \mathbb{E}[t(\tilde{e}_N)|\xi(n)]$$

and the usual Bayes estimator is given by

$$\mathbb{E}[t(\tilde{p})|\xi(n)].$$

Example 1 (Estimation of the mean)**.** *Suppose the statistician has to estimate the mean under the squared error loss, i.e. the functional of interest is* $t(p) := \int_{\mathbb{R}} xp(dx)$*. The usual Bayes estimator is*

$$\hat{\mu}_n := \mathbb{E}[\xi_{n+1}|\xi(n)]$$

while the "finitary Bayes" estimator is

$$\hat{\mu}_{FB} = \frac{n}{N}\bar{\mu}_n + \frac{N-n}{N}\hat{\mu}_n,$$

where

$$\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i.$$

Note that in this case the finitary Bayes estimator is a convex combination of the usual Bayes estimator with the empirical (plug–in) estimator $\bar{\mu}_n$*.*

Example 2 (Estimation of the variance)**.** *Now, consider the estimation of the variance* $t(p) = \int_{\mathbb{R}} x^2 p(dx) - (\int_{\mathbb{R}} xp(dx))^2$*, under the squared error loss. In this case the space of actions is* $\mathbb{R}^+$ *and the usual Bayes estimator is*

$$\hat{\sigma}^2_n := \hat{s}^2_n - \hat{c}_{1,2,n}$$

where

$$\hat{s}^2_n := E[\xi^2_{n+1}|\xi(n)] \qquad \textit{and} \qquad \hat{c}_{1,2,n} := E[\xi_{n+1}\xi_{n+2}|\xi(n)].$$

Some computations show that the "finitary Bayes" estimator is

$$\hat{\sigma}^2_{FB} = \frac{n}{N}\bar{s}^2_n + \frac{N-n+n/N-1}{N}\hat{s}^2_n - \frac{n^2}{N^2}\bar{c}_{1,2,n} - \frac{(N-n)(N_n-1)}{N^2}\hat{c}_{1,2,n} - \frac{2(N-n)n}{N^2}\bar{\mu}_n\hat{\mu}_n$$

where

$$\bar{s}^2_n := \frac{1}{n}\sum_{i=1}^{n}\xi_i^2 \qquad \textit{and} \qquad \bar{c}_{1,2,n} := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j.$$

Example 3 (Estimation of the distribution function)**.** *Assume one has to estimate* $t(p) = F_p(y) = p\{(-\infty, y]\}$*, where* y *is a fixed real number. Under the square loss function, the classical Bayes estimator is*

$$\mathbb{E}(\mathbb{I}_{(-\infty,y]}(\xi_{n+1})|\xi(n))$$

while the "finitary Bayes" estimator is

$$\hat{F}_{FB}(y) = \frac{n}{N}E_n(y) + \frac{N-n}{N}\mathbb{E}(\mathbb{I}_{(-\infty,y]}(\xi_{n+1})|\xi(n))$$

where $E_n(y) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{(-\infty,y]}(\xi_i)$*.*

Example 4 (Estimation of the mean difference)**.** *Estimate the Gini mean difference*

$$t(p) = \Delta(p) = \int_{\mathbb{R}^2} |x-y|p(dx)p(dy)$$

under the squared error loss. The usual Bayes estimator is

$$\mathbb{E}(|\xi_{n+1} - \xi_{n+2}||\xi(n))$$

while the "finitary Bayes" estimator is

$$\begin{aligned}\mathbb{E}(\Delta(\tilde{e}_N)|\xi(n)) = \frac{n^2}{N^2}\Delta_n + \frac{(N-n)^2-(N-n)}{N^2}\mathbb{E}(|\xi_{n+1} - \xi_{n+2}||\xi(n)) \\ + \frac{2(N-n)}{N^2}\sum_{j<n}\mathbb{E}(|\xi_j - \xi_{n+1}||\xi(n)),\end{aligned}$$

where

$$\Delta_n := \frac{1}{n^2}\sum_{i,j\le n}|\xi_i - \xi_j|.$$

It is worth noticing that in all the previous examples when N goes to $+\infty$ the "finitary Bayes" estimator converges to the usual Bayes estimator, while the "finitary Bayes" estimator becomes the usual plug–in frequentistic estimator if $n = N$.

3. Comparison between posterior distributions of means

Let Q be the probability distribution of $\tilde{p}$. Then Q turns out to be a probability measure on $\mathbb{P}(X)$. Without loss of generality consider $\mathbb{P} = \mathbb{P}(X)$ endowed with a bounded metric λ which induces the weak convergence on $\mathbb{P}$ (e.g. the Prohorov metric), and set $\mathcal{P}$ for its Borel σ–field. In what follows, if necessary, expand $(\Omega, \mathcal{F}, P)$ in order to contain all the random variables needed and, for any random variable V, let $\mathcal{L}_V$ designate the probability distribution of V and, for any other random element U, by $\mathcal{L}_{V|U}$ denote some conditional probability distribution for V given U. In particular, $\mathcal{L}_{\tilde{e}_N|\xi(n)}$ will denote (a version of) the conditional distribution of $\tilde{e}_N$ given $\xi(n) := (\xi_1, \dots, \xi_n)$ and $\mathcal{L}_{\tilde{p}|\xi(n)}$ will stand for (a version of) the conditional distribution of $\tilde{p}$ given $\xi(n)$, i.e. the so–called posterior distribution of $\tilde{p}$. Such distributions exist since $(\mathbb{P}, \lambda)$ is Polish.

As already said, the main goal of this paper is comparing $\mathcal{L}_{\tilde{e}_N|\xi(n)}$ with $\mathcal{L}_{\tilde{p}|\xi(n)}$. We start by comparing posterior means. Indeed, as we have seen in the previous section, the posterior mean of a function f appears in many natural statistical estimation problems. For the sake of notational simplicity, for any measurable real–valued function f, set

$$\tilde{e}_N(f) := \int_X f(x)\tilde{e}_N(dx) = \frac{1}{N}\sum_{i=1}^N f(\xi_i)$$

and

$$\tilde{p}\,(f) := \int_X f(x)\,\tilde{p}\,(dx).$$

First of all we prove this very simple

Proposition 3.1. *Given a real–valued measurable function f, if $P\{\tilde{p}(|f|) < +\infty\} = 1$, then $\tilde{e}_N(f)$ converges in law to $\tilde{p}(f)$ (as $N \to +\infty$). Analogously, $\mathcal{L}_{\tilde{e}_N(f)|\xi(n)}$ converges weakly (almost surely) to $\mathcal{L}_{\tilde{p}(f)|\xi(n)}$.*

Proof. Let ϕ be a bounded continuous function with $c = \|\phi\|_\infty$. Then $\phi(\tilde{e}_N) \leq c$. Now, $\mathbb{E}(\phi(\tilde{e}_N(f)|\tilde{p})$ converges almost surely to $\mathbb{E}(\phi(\tilde{p}(f))|\tilde{p})$. To see this, note that, conditionally on $\tilde{p}$, $\tilde{e}_N(f)$ is a sum of independent random variables with mean $\tilde{p}(f)$ and absolute moment $\tilde{p}(|f|)$, and, since $\tilde{p}(|f|)$ is almost surely finite, the conditional law of $\tilde{e}_N(f)$ given $\tilde{p}$ converges almost surely to $\tilde{p}(f)$, and hence also in law. Since $|\mathbb{E}(\phi(\tilde{e}_N(f)|\tilde{p})| \leq c$ almost surely, to conclude the proof it is enough to apply the dominated convergence theorem. The second part of the theorem can be proved in the same way conditioning with respect to $(\tilde{p}, \xi(n))$. ♢

In order to give a quantitative version of the previous statement we resort to the so–called Gini–Kantorovich–Wasserstein distance. Let $\mathbb{P}_1 = \mathbb{P}_1(\mathbb{R}^d)$ be the subset of the set $\mathbb{P} = \mathbb{P}(\mathbb{R}^d)$ of all probability measures on $\mathcal{B}(\mathbb{R}^d)$ defined by $\mathbb{P}_1 := \{p \in \mathbb{P} : \int_{\mathbb{R}^d} \|x\| p(dx) < +\infty\}$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. For every couple of probability measures (p, q) in $\mathbb{P}_1 \times \mathbb{P}_1$ the Gini-Kantorovich-Wasserstein distance (of order one) between p and q is defined by

$$w_1(p, q) := \inf\left\{\int_{\mathbb{R}^{2d}} \|x - y\|\gamma(dxdy) : \gamma \in \mathcal{M}(p, q)\right\},$$

$\mathcal{M}(p, q)$ being the class of all probability measures on $(\mathbb{R}^{2\times d}, \mathcal{B}(\mathbb{R}^{2\times d}))$ with marginal distributions p and q. For a general definition of the Gini-Kantorovich-Wasserstein distance and its properties see, e.g., [30]. If Z_1 and Z_2 are two random variables with law p and q respectively, $w_1(Z_1, Z_2)$ will stand for $w_1(p, q)$.

Proposition 3.2. *Given a real–valued measurable function f such that $\mathbb{E}[f(\xi_1)^2] < \infty$, then*

$$w_1(\tilde{e}_N(f), \tilde{p}(f)) \leq \frac{1}{\sqrt{N}}\left(\mathbb{E}|f(\xi_1) - \tilde{p}\,(f)\,|^2\right)^{1/2} \leq \frac{2}{\sqrt{N}}\sqrt{\mathbb{E}[f(\xi_1)^2]}.$$

Moreover,

$$w_1(\mathcal{L}_{\tilde{e}_N(f)|\xi(n)}, \mathcal{L}_{\tilde{p}(f)|\xi(n)}) \leq \frac{n}{N}\left(\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) + \mathbb{E}[\tilde{p}\,(f)|\,\xi(n)]\right) + \frac{2}{\sqrt{N-n}}\left(\mathbb{E}[f(\xi_{n+1})^2|\xi(n)]\right)^{1/2} \qquad (a.e.).$$

Proof. Applying a well–known conditioning argument, note that

$$\begin{aligned} w_1(\tilde{e}_N(f),\tilde{p}(f)) &\leq \mathbb{E}|\tilde{e}_N(f)-\tilde{p}(f)| \\ &= \mathbb{E}[\mathbb{E}(|\tilde{e}_N(f)-\tilde{p}(f)||\tilde{p})] = \mathbb{E}\Big[\mathbb{E}\Big(\Big|\frac{1}{N}\sum_{i=1}^N f(\xi_i)-\int f\tilde{p}(dx)\Big|\Big|\tilde{p}\Big)\Big] \\ &\quad \text{(by the Cauchy-Schwartz inequality)} \\ &\leq \frac{1}{\sqrt{N}}\mathbb{E}\left[\mathbb{E}\Big[\Big(f(\xi_1)-\int f\tilde{p}\Big)^2\Big|\tilde{p}\Big]^{1/2}\right] \\ &\quad \text{(by the Jensen inequality)} \\ &\leq \frac{1}{\sqrt{N}}\mathbb{E}\left[(f(\xi_1)-\int f\tilde{p})^2\right]^{1/2}. \end{aligned}$$

Clearly, $\mathbb{E}[(f(\xi_1)-\int f\tilde{p})^2]^{1/2} \leq (2(\mathbb{E}[f(\xi_1)^2+(\int f\tilde{p})^2)^{1/2}$ and, by the Jensen inequality, $\mathbb{E}(\int f\tilde{p})^2 \leq \mathbb{E}(\int f^2\tilde{p}) = \mathbb{E}[f(\xi_1)^2]$. As for the second part of the proposition, first note that

$$\begin{aligned} w_1(\mathcal{L}_{\tilde{p}(f)|\xi(n)},\mathcal{L}_{\tilde{e}_N(f)|\xi(n)}) \leq \frac{N-n}{N}\mathbb{E}\left[\frac{1}{N-n}\sum_{i=n+1}^N f(\xi_i)-\tilde{p}\,(f)\,\Big|\xi(n)\right] \\ + \frac{n}{N}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(\xi_i)-\tilde{p}\,(f)\,\Big|\xi(n)\right]. \end{aligned}$$

Now, take the conditional expectation given $(\tilde{p},\xi(n))$ and use again the Cauchy-Schwartz inequality to obtain

$$\begin{aligned} &\mathbb{E}[\frac{1}{N-n}\sum_{i=n+1}^N f(\xi_i)-\tilde{p}\,(f)\,|\xi(n)] \\ &\qquad \leq \frac{1}{\sqrt{N-n}}\mathbb{E}[\mathbb{E}\big(|f(\xi_{n+1})-\tilde{p}\,(f)|^2\,|\tilde{p},\xi(n)\big)\big|\xi(n)]^{1/2}. \end{aligned}$$

Finally, to complete the proof, apply the Jensen inequality and argue as in the previous part of the proof. ♢

Of course, the mean is not the unique interesting functional which appears in statistical problems. For instance, statisticians frequently deal with functionals of the form

$$t_1(p)=\int_{X^k} f(x_1,\dots,x_k)p(dx_1)\dots p(dx_k),$$

or even of the form

$$t_2(p)=\mathrm{argmin}_{\theta\in\Theta}\int_{X^k} f_\theta(x_1,\dots,x_k)p(dx_1)\dots p(dx_k).$$

Think, for example, of the variance or of the median of a probability measure, respectively.

It is immediate to generalize Proposition 3.1 according to

Proposition 3.3. *Given a measurable function $f: X^k \to \mathbb{R}$ such that*

$$P\{\int_{X^k}|f(x_1,\dots,x_k)|\tilde{p}(dx_1)\dots\tilde{p}(dx_k)<+\infty\}=1,$$

then $t_1(\tilde{e}_N)$ converges in law to $t_1(\tilde{p})$ (as $N\to+\infty$). Analogously, $\mathcal{L}_{\xi(n)|t_1(\tilde{e}_N)}$ converges weakly (almost surely) to $\mathcal{L}_{\xi(n)|t_1(\tilde{p})}$.

As far as functionals of the type of t_2 are concerned, the situation is less simple. As a general strategy, one could apply the usual *argmax argument*. See, e.g., [33]. To do this, set

$$\mathbb{M}_N(\theta) := \int_{X^k} f_\theta(x_1, \dots, x_k)\tilde{e}_N(dx_1)\dots\tilde{e}_N(dx_k),$$

$$\mathbb{M}(\theta) := \int_{X^k} f_\theta(x_1, \dots, x_k)\tilde{p}(dx_1)\dots\tilde{p}(dx_k)$$

and $\theta_N = t_2(\tilde{e}_N)$. Assume that Θ is a subset of $\mathbb{R}^d$ and, for every $T \subset \mathbb{R}^d$ define the set $l^\infty(T)$ of all measurable functions $f : T \to \mathbb{R}$ satisfying

$$\|f\|_T := \sup_{t\in T} |f(t)| < +\infty.$$

A version of the argmax theorem (Theorem 3.2.2 in [33]) implies that: *If $\mathbb{M}_N$ converges in law to $\mathbb{M}$ in $l^\infty(K)$ for every compact set $K \subset \mathbb{R}^d$, if almost all sample paths $\theta \mapsto \mathbb{M}(\theta)$ are lower semi-continuous and possess a unique minimum at a random point $\hat{\theta} = t_2(\tilde{p})$, and if $(\theta_N)_{N\geq 1}$ is tight, then θ_N converges in law to $\hat{\theta}$.*

As for the first hypothesis, that is $\mathbb{M}_N$ converges in law to $\mathbb{M}$ in $l^\infty(K)$ for every compact set $K \subset \mathbb{R}^d$, one can resort to Theorems 1.5.4 and 1.5.6 in [33]. Such theorems imply that if $(\mathbb{M}_N(\theta_1), \dots, \mathbb{M}_N(\theta_k))$ converges in law to $(\mathbb{M}(\theta_1), \dots, \mathbb{M}(\theta_k))$ for every k and every $(\theta_1, \dots, \theta_k)$ in K^k and if, for every ϵ and $\eta > 0$, there is a finite partition $\{T_1, \dots, T_N\}$ of K such that

$$\limsup_N P\{\sup_i \sup_{h_1,h_2\in T_i} |\mathbb{M}_N(h_1) - \mathbb{M}_N(h_2)| > \epsilon\} < \eta \tag{1}$$

then $\mathbb{M}_N$ converges in law to $\mathbb{M}$ in $l^\infty(K)$ for every compact set $K \subset \mathbb{R}^d$. Hence, one can try to show that

$$|f_{\theta_1}(x_1, \dots, x_k) - f_{\theta_2}(x_1, \dots, x_k)| \leq g(\|\theta_1 - \theta_2\|_2)\phi(x_1, \dots, x_k)$$

for some continuous function g, with $g(0) = 0$, and some function ϕ such that for some θ_0

$$P\left\{\int_{X^k} [\phi(x_1, \dots, x_k) + |f_{\theta_0}(x_1, \dots, x_k)|]\tilde{p}(dx_1)\dots\tilde{p}(dx_k) < +\infty\right\} = 1.$$

If these conditions hold, then the convergence of $\mathbb{M}_N$ to $\mathbb{M}$ is easily proved, whereas both tightness of $(\theta_N)_{N\geq 1}$ and uniqueness of $\hat{\theta}$ require additional assumptions.

Here is an example, where $med(p)$ denotes the median of the distribution p.

Proposition 3.4. *Let $M_N = med(\tilde{e}_{2N+1})$, that is $M_N = \xi_{(N+1)}$ if $\xi_{(1)} \leq \dots \leq \xi_{(2N+1)}$. If*

$$P\left\{\int_{\mathbb{R}} |x|\tilde{p}(dx) < +\infty\right\} = 1 \qquad \text{and} \qquad P\{med(\tilde{p}) \text{ is unique}\} = 1,$$

then M_N converges in law to $med(\tilde{p})$ as N diverges. Analogously, if, for some $n < N$,

$$\xi(n) \mapsto P\left\{\int_{\mathbb{R}} |x|\tilde{p}(dx) < +\infty \;\middle|\; \xi(n)\right\} = 1 \qquad (a.e.)$$

and

$$\xi(n) \mapsto P\{med(\tilde{p}) \text{ is unique} \,|\xi(n)\} = 1, \qquad (a.e.),$$

then $\mathcal{L}_{M_N|\xi(n)}$ converges weakly (almost surely) to $\mathcal{L}_{med(\tilde{p})|\xi(n)}$ as N diverges.

Proof. In this case

$$\mathbb{M}_N(\theta) = \int_{\mathbb{R}} |x-\theta| d\tilde{e}_{2N+1}$$

and

$$\mathbb{M}(h) = \int_{\mathbb{R}} |x-\theta| d\tilde{p}.$$

Since $P\left\{\int_{\mathbb{R}} |x| \tilde{p}(dx) < +\infty\right\} = 1$, from Proposition 3.3 we get that $(\mathbb{M}_N(\theta_1), \dots, \mathbb{M}_N(\theta_k))$ converges in law to $(\mathbb{M}(\theta_1), \dots, \mathbb{M}(\theta_k))$ for every k and every $(\theta_1, \dots, \theta_k)$. Moreover

$$|\mathbb{M}_N(\theta_1) - \mathbb{M}_N(\theta_2)| \leq |\theta_1 - \theta_2|,$$

hence (1) is verified. It remains to prove the tightness of $(M_N)_{N\geq 1}$. First of all observe that if $X_1, \dots, X_{2N+1}$ are i.i.d random variables with common distribution function F then the distribution function of the median of $X_1, \dots, X_{2N+1}$ is given by

$$x \mapsto \sum_{k=N+1}^{2N+1} \binom{2n+1}{k} F^k(x)(1-F(x))^{2N+1-k} = \frac{1}{B(N+1,N+1)} \int_0^{F(x)} t^N (1-t)^N dt,$$

where B is the Euler integral of the first kind (the so–called beta function). Hence, denoting by $\tilde{F}(x)$ the distribution function of $\tilde{p}$ and setting

$$H_x(\tau) = P\{\tilde{F}(x) \leq \tau\},$$

it follows that

$$\begin{aligned} P\{M_N \leq x\} &= \mathbb{E}\left(B(N+1,N+1)^{-1} \int_0^{\tilde{F}(x)} t^N (1-t)^N dt \right) \\ &= B(N+1,N+1)^{-1} \int_0^1 \int_0^\tau t^N (1-t)^N dt dH_x(\tau) \\ &= B(N+1,N+1)^{-1} \int_0^1 t^N (1-t)^N [1-H_x(t)] dt. \end{aligned}$$

Now, by the Markov inequality,

$$[1-H_x(t)] = P\{\tilde{F}(x) > t\} \leq \frac{1}{t}\mathbb{E}[\tilde{F}(x)] = P\{\xi_1 \leq x\},$$

hence,

$$P\{M_N \leq x\} \leq P\{\xi_1 \leq x\} B(N+1,N+1)^{-1} \int_0^1 t^N (1-t)^N dt = P\{\xi_1 \leq x\} \frac{2N+1}{N}.$$

In the same way, it is easy to see that

$$\begin{aligned} P\{M_N < x\} &= 1 - P\{M_N \leq x\} = 1 - B(N+1,N+1)^{-1} \int_0^1 t^N (1-t)^N [1-H_x(t)] dt \\ &= B(N+1,N+1)^{-1} \int_0^1 t^N (1-t)^N H_x(t) dt \\ &= B(N+1,N+1)^{-1} \int_0^1 t^N (1-t)^N H_x(1-t) dt \end{aligned}$$

and hence

$$P\{M_N > x\} \leq \frac{2N+1}{N} P\{\xi_1 \geq x\}.$$

With these inequalities it is immediate to prove the tightness of $(M_N)_{N\geq 1}$. The proof of the second part of the proposition is analogous. $\diamondsuit$

4. Comparing posterior distributions of random probabilities

We now turn our attention to the comparison of $\mathcal{L}_{\tilde{e}_N|\xi(n)}$ with $\mathcal{L}_{\tilde{p}|\xi(n)}$. We shall use the Gini-Kantorovich-Wasserstein distance on the space of all probability measures. The Gini-Kantorovich-Wasserstein distance of order 1 (relative to a metric λ) between two probability measures, say (Q_1, Q_2), defined on $(\mathbb{P}, \mathcal{P})$ is

$$W_1(Q_1, Q_2) := \inf\left\{\int_{\mathbb{P}^2} \lambda(p_1, p_2)\Gamma(dp_1 dp_2) : \Gamma \in M(Q_1, Q_2)\right\}$$

where $M(Q_1, Q_2)$ is the set of all probability measures on $(\mathbb{P}\times\mathbb{P}, \mathcal{P}\otimes\mathcal{P})$ with marginals Q_1 and Q_2. Here, it is worth recalling that W_1 admits the following dual representation

$$\begin{aligned} W_1(Q_1, Q_2) = \sup\Big\{ &\int_{\mathbb{P}} f(p)(Q_1(dp) - Q_2(dp)); \\ &f : \mathbb{P}\to\mathbb{R}, \ \ |f(p) - f(q)| \leq \lambda(p, q) \ \ \forall\, p, q \ \in \ \mathbb{P}\Big\}. \end{aligned} \tag{2}$$

See, e.g., Theorem 11.8.2 in [9]. The main goal of this section is to give explicit upper bounds for the random variable $W_1(\mathcal{L}_{\tilde{e}_N|\xi(n)}, \mathcal{L}_{\tilde{p}|\xi(n)})$.

4.1. A first bound for the posterior distributions.

There is a large body of literature on the rate of convergence to zero (when N diverges) of

$$E_N(p) := \mathbb{E}\left[\lambda\left(p, \nu_N^{(p)}\right)\right],$$

where $\nu_N^{(p)} := \sum_{i=1}^N \delta_{z_i^{(p)}}/N$ and $(z_i^{(p)})_{i\geq 1}$ is a sequence of independent and identically distributed (i.i.d.) random variables taking values in X, with common probability measure p. See, for instance, [1], [8] and [18]. The next lemma shows how these well–known results can be used to get a bound for $W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)})$.

Lemma 4.1. *Assume that λ is bounded and satisfies*

$$\lambda(p, \epsilon p_1 + (1-\epsilon)p_2) \leq \epsilon\lambda(p, p_1) + (1-\epsilon)\lambda(p, p_2) \tag{3}$$

for every ϵ in $(0,1)$ and every p, p_1, p_2 in $\mathbb{P}$. Moreover, let $K := \sup\{\lambda(p,q) : (p,q) \in \mathbb{P}^2\}$. Then

$$W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) \leq \int_{\mathbb{P}} E_{N-n}(p)\mathcal{L}_{\tilde{p}|\xi(n)}(dp) + \frac{nK}{N} \tag{4}$$

holds true for P–almost every $\xi(n)$.

Proof. First of all, note that for every A in $\mathcal{P}$

$$\mathcal{L}_{\tilde{e}_N|\xi(n)}(A) = \int_{\mathbb{P}} \mathcal{L}_{\tilde{e}_N|\xi(n),\tilde{p}}(A)\mathcal{L}_{\tilde{p}|\xi(n)}(dp)$$

where , according to our notation, $\mathcal{L}_{\tilde{e}_N|\xi(n),\tilde{p}}$ denotes (a version of) the conditional distribution of $\tilde{e}_N$ given $(\xi(n), \tilde{p})$. Hence, from the dual representation (2) of W_1 it is easy to see that

$$W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) \leq \int_{\mathbb{P}} W_1\left(\delta_p, \mathcal{L}_{\tilde{e}_N|\xi(n),\tilde{p}}\right)\mathcal{L}_{\tilde{p}|\xi(n)}(dp).$$

Now, write

$$\tilde{e}_N = \frac{n}{N}\tilde{e}_n + \frac{N-n}{N}\tilde{e}_{N,n}$$

with $\tilde{e}_{N,n} = \sum_{i=n+1}^{N} \delta_{\xi_i}/(N-n)$, and observe that $\tilde{e}_n$ and $\tilde{e}_{N,n}$ are conditionally independent given $\tilde{p}$. Moreover, $\tilde{e}_{N,n}$ has the same law of $\tilde{e}_{N-n}$ and $W_1(\delta_p, Q) = \int_{\mathbb{P}} \lambda(p,q) Q(dq)$. Hence,

$$\begin{aligned} W_1\left(\delta_p, \mathcal{L}_{\tilde{e}_N|\xi(n),\tilde{p}}\right) &= \int_{\mathbb{P}} \lambda(p,q) \mathcal{L}_{\tilde{e}_N|\xi(n),\tilde{p}}(dq) \\ &= \mathbb{E}\left[\lambda\Big(p, \frac{1}{N}\sum_{i=1}^{N-n} \delta_{z_i^{(p)}} + \frac{1}{N}\sum_{i=1}^{n} \delta_{\xi_i}\Big)\right] \\ &\leq \frac{N-n}{N}\mathbb{E}\left[\lambda\Big(p, \frac{1}{N-n}\sum_{i=1}^{N-n} \delta_{z_i^{(p)}}\Big)\right] + \frac{nK}{N} = \frac{N-n}{N} E_{N-n}(p) + \frac{nK}{N}. \end{aligned}$$

The thesis follows from integration over $\mathbb{P}$ with respect to $\mathcal{L}_{\tilde{p}|\xi(n)}$. $\diamondsuit$

In the next three subsections we shall use the previous lemma with different choices of X and λ.

4.2. **The finite case.** We start from the simple case in which $X = \{a_1, \dots, a_k\}$. Here $\mathbb{P}$ can be seen as the simplex

$$\mathcal{S}_k = \{x \in \mathbb{R}^k : 0 \leq x_i \leq 1, i = 1 \dots, k, \sum_{i=1}^{k} x_i = 1\}.$$

Define λ to be the total variation distance, i.e. $\lambda(p,q) = \frac{1}{2}\sum_{i=1}^{k} |p(a_i) - q(a_i)|$. In point of fact, it should be noted that, since X is finite, there is no difference between the strong and the weak topology on $\mathbb{P}$. In this case, for every $j = 1, \dots, k$, one has

$$\tilde{e}_N(a_j) = \sharp\{i : \xi_i = a_j;\ 1 \leq i \leq N\}/N.$$

Now, denoting by Z_i a binomial random variable of parameters $(N-n, p_i)$ ($p_i := p(a_i)$), we get

$$\begin{aligned} \mathbb{E}\left[\lambda\left(p, \nu_{N-n}^{(p)}\right)\right] &= \frac{1}{2(N-n)}\sum_{i=1}^{k} \mathbb{E}[|Z_i - (N-n)p_i|] \\ &\leq \frac{1}{2(N-n)}\sum_{i=1}^{k} \sqrt{\mathbb{E}[|Z_i - (N-n)p_i|^2]} \\ &= \frac{1}{2(N-n)}\sum_{i=1}^{k} \sqrt{(N-n)p_i(1-p_i)} = \frac{1}{2\sqrt{N-n}}\sum_{i=1}^{k} \sqrt{p_i(1-p_i)} \\ &\leq \frac{k}{4\sqrt{(N-n)}}. \end{aligned}$$

Observing that $K := \sup\{TV(p,q) : p, q \in \mathcal{S}_k\} \leq 1$ and that the total variation distance satisfies (3), Lemma 4.1 gives

Proposition 4.2. *If $X = \{a_1, \dots, a_k\}$, then*

$$W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) \leq \frac{k}{4\sqrt{N-n}} + \frac{n}{N}.$$

4.3. **The case $X = \mathbb{R}$.** Passing to a general Euclidean space we first need to choose a suitable metric λ. We recall that if p and q belongs to $\mathbb{P}(\mathbb{R}^d)$, the so–called bounded Lipschitz distance (denoted by β) between p and q is defined by

$$\beta(p,q) = \sup\left\{ \int_{\mathbb{R}^d} f(x)[p(dx) - q(dx)];\ f : \mathbb{R}^d \to \mathbb{R},\ \|f\|_{BL} \leq 1 \right\}$$

where $\|f\|_{BL} := \sup_{x\in\mathbb{R}^d} |f(x)| + \inf_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d} |f(x) - f(y)|/\|x - y\|$. See Section 11.3 in [9]. Note that $\sup_{(p,q)\in\mathbb{P}} \beta(p,q) \leq 2$ and that β satisfies $\beta(p, \epsilon p_1 + (1-\epsilon)p_2) \leq \epsilon\beta(p,p_1) + (1-\epsilon)\beta(p,p_2)$ for every ϵ in $(0,1)$ and every p, p_1, p_2 in $\mathbb{P}$. Recall also that β metrizes the weak topology (see, e.g., Theorem 11.3.3 in [9]). In what follows we take $X = \mathbb{R}$ and $\lambda = \beta$. As a consequence of Lemma 4.1, we get the next proposition in which, for every p in $\mathbb{P}$, we set $F_p(x) = p\{(-\infty, x]\}$.

Proposition 4.3. *Let $X = \mathbb{R}$ and $\lambda = \beta$. Set $\Delta(p) := \int_{\mathbb{R}} \sqrt{F_p(t)(1-F_p(t))}dt$. If $\mathbb{E}[\Delta(\tilde{p})] < +\infty$, then the inequalities*

$$\begin{aligned} W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) &\leq \frac{1}{\sqrt{N-n}}\ \mathbb{E}[\Delta(\tilde{p})|\xi(n)] + \frac{2n}{N} \\ &\leq \frac{1}{\sqrt{N-n}} Y + \frac{2n}{N}, \end{aligned}$$

holds true for all $n < N$, with $Y := \sup_n \mathbb{E}[\Delta(\tilde{p})|\xi(n)] < +\infty$, for P–almost every $\xi(n)$.

Proof. As already recalled, $\sup_{(p,q)\in\mathbb{P}^2} \beta(p,q) \leq 2$ and β satisfies (3). Using the dual representation of w_1– which is the analogue of (2) with $\mathbb{R}^d$ in the place of $\mathbb{P}$ and $\|\cdot\|$ in the place of λ– it is easy to see that

$$\beta(p,q) \leq w_1(p,q). \tag{5}$$

Moreover, recall that, when $X = \mathbb{R}$,

$$w_1(p,q) = \int_{\mathbb{R}} |F_p(x) - F_q(x)|dx. \tag{6}$$

See, for instance, [30]. For any p in $\mathbb{P}_1$, for the sake of simplicity, set $z_i^{(p)} = z_i$ and observe that combination of (5) and (6) gives

$$\begin{aligned} \mathbb{E}\left[\lambda\Big(p, \frac{1}{N-n}\sum_{i=1}^{N-n}\delta_{z_i}\Big)\right] &\leq \mathbb{E}\left[\int_{\mathbb{R}} |F_p(t) - \frac{1}{N-n}\sum_{i=1}^{N-n}\mathbb{I}(z_i \leq t)|dt\right] \\ &= \frac{1}{N-n}\mathbb{E}\left[\int_{\mathbb{R}} |(N-n)F_p(t) - \sum_{i=1}^{N-n}\mathbb{I}(z_i \leq t)|dt\right] \\ &\qquad \text{(by Fubini theorem)} \\ &= \frac{1}{N-n}\int_{\mathbb{R}} \mathbb{E}\left[|(N-n)F_p(t) - \sum_{i=1}^{N-n}\mathbb{I}(z_i \leq t)|\right] dt. \end{aligned}$$

Now, note that $\sum_{i=1}^{N-n}\mathbb{I}(z_i \leq t)$ are binomial random variables of parameters $((N-n), F_p(t))$. Hence, since

$$\int_{\mathbb{R}} \sqrt{F_p(t)(1-F_p(t))}\, dt < +\infty \tag{7}$$

holds true P-almost surely, from the Cauchy-Schwartz inequality one gets

$$\mathbb{E}\left[\lambda\Big(p,\nu_{N-n}^{(p)}\Big)\right] \leq \frac{1}{\sqrt{N-n}}\int_{\mathbb{R}}\sqrt{F_p(t)(1-F_p(t))}dt.$$

Combination of this fact with Lemma 4.1 and the obvious identity $\int_{\mathbb{P}}\Delta(p)\mathcal{L}_{\tilde{p}|\xi(n)}(dp) = \mathbb{E}[\Delta(\tilde{p})|\xi(n)]$ gives the first part of the thesis. To conclude the proof, apply Doob's martingale convergence theorem (see, e.g., Theorem 10.5.1 in [9]) to $\mathbb{E}[\Delta(\tilde{p})|\xi(n)]$ in order to prove that $\sup_n \mathbb{E}[\Delta(\tilde{p})|\xi(n)] < +\infty$ almost surely. $\diamondsuit$

A first simple consequence of the previous proposition is embodied in

Corollary 1. *Let $X = [-M, M]$ for some $0 < M < +\infty$. Then,*

$$W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) \leq \frac{2M}{\sqrt{N-n}} + \frac{2n}{N}$$

holds true for all $n < N$ for P–almost every $\xi(n)$.

It is worth recalling that $\Delta(p) < +\infty$ implies finite second moment for p but not conversely (this condition defines the Banach space $L_{2,1}$, cf. [22], p.10). It is easy to show that if p has finite moment of order $2+\delta$, for some positive δ, then

$$\Delta(p) \leq \left[1 + C_\delta\left(\int_{\mathbb{R}}|x|^{2+\delta}p(dx)\right)^{1/2}\right] \tag{8}$$

holds true with $C_\delta := \sqrt{2(1+\delta)/\delta}$. As a consequence of these statements we have the following

Corollary 2. *If $\mathbb{E}[\Delta(\tilde{p})] < +\infty$, then*

$$\mathbb{E}[W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)})] \leq \frac{\mathbb{E}[\Delta(\tilde{p})]}{\sqrt{N-n}} + \frac{2n}{N}$$

and

$$P\left\{W_1(\mathcal{L}_{\tilde{e}_N|\xi(n)}, \mathcal{L}_{\tilde{p}|\xi(n)}) > \epsilon\right\} \leq \frac{1}{\epsilon}\left[\frac{\mathbb{E}[\Delta(\tilde{p})]}{\sqrt{N-n}} + \frac{2n}{N}\right]$$

hold true for all $n < N$. Moreover, if $\mathbb{E}|\xi_1|^{2+\delta} < +\infty$ for some positive δ, then

$$\mathbb{E}[\Delta(\tilde{p})] \leq \left[1 + \left(\frac{2(1+\delta)}{\delta}\mathbb{E}|\xi_1|^{2+\delta}\right)^{1/2}\right].$$

Proof. By Proposition 4.3, whenever $\mathbb{E}[\Delta(\tilde{p})] < +\infty$, one can write

$$\begin{aligned}\mathbb{E}[W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)})] &\leq \frac{1}{\sqrt{N-n}}\mathbb{E}\left[\mathbb{E}\big[\Delta(\tilde{p})|\xi(n)\big]\right] + \frac{2n}{N} \\ &= \frac{\mathbb{E}[\Delta(\tilde{p})]}{\sqrt{N-n}} + \frac{2n}{N}.\end{aligned}$$

Now, let $\bar{p}(\cdot) = \mathbb{E}(\tilde{p}(\cdot))$. Then, (8) together with Fubini theorem and Jensen inequality yield

$$\mathbb{E}[\Delta(\tilde{p})] \leq \left[1 + C_\delta\left(\int_{\mathbb{R}}|x|^{2+\delta}\bar{p}(dx)\right)^{1/2}\right].$$

Combining these facts with Markov inequality completes the proof. $\diamondsuit$

4.4. **The case $X = \mathbb{R}^d$.** Let $X = \mathbb{R}^d$ and $\lambda = \beta$. For any p in $\mathbb{P}$ and k in $\mathbb{N}$ consider

$$\Psi_k(p) := \left(\sup_{\epsilon \in (0,1]} \epsilon^k N(\epsilon, \epsilon^{k/(k-2)}, p) \right)^{1/2}$$

where $N(\epsilon, \eta, p)$ is the minimal number of sets of diameter $\leq 2\epsilon$ which cover $\mathbb{R}^d$ except for a set A with $p(A) \leq \eta$. Proposition 3.1 in [8] (see also Theorem 7 in [18]) gives

$$\mathbb{E}\left[\beta\big(\, p\,,\, \nu_{N-n}^{(p)}\,\big)\right] \leq (N-n)^{-1/k}[\frac{4}{3} + 4 \cdot 3^{2k}\Psi_k(p)].$$

Using the last inequality and arguing as in the proof of Proposition 4.3 we obtain the following

Proposition 4.4. *If $\mathbb{E}[\Psi_k(\tilde{p})] < +\infty$ for some positive k, then the inequality*

$$\begin{aligned} W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) &\leq \frac{1}{(N-n)^{1/k}} \left(\frac{4}{3} + 4 \cdot 3^{2k} \mathbb{E}[\Psi_k(\tilde{p})|\xi(n)] \right) + \frac{2n}{N} \\ &\leq \frac{1}{(N-n)^{1/k}} \left(\frac{4}{3} + 4 \cdot 3^{2k} Y \right) + \frac{2n}{N}, \end{aligned}$$

holds true for all $n < N$, with $Y := \sup_n \mathbb{E}[\Psi_k(\tilde{p})|\xi(n)] < +\infty$, for P–almost every $\xi(n)$.

Remark 1. In the last proposition the fact that $X = \mathbb{R}^d$ does not play any special role. Everything remains true if X is a Polish space.

Condition $\mathbb{E}[\Psi_k(\tilde{p})] < +\infty$ is almost impossible to check. In what follows we will assume a more tractable hypothesis. If $\int_{\mathbb{R}^d} \|x\|^\gamma p(dx) < +\infty$ where $\gamma = \frac{kd}{(k-d)(k-2)}$, $d \geq 2$ and $k > d$, Proposition 3.4 in [8] (see also Theorem 8 in [18]) yields

$$\Psi_k(p)^2 \leq 2^d \left[1 + 2 \left(\int_{\mathbb{R}^d} \|x\|^\gamma p(dx) \right)^{1/\gamma} \right]. \tag{9}$$

Using this last inequality we can prove the following

Proposition 4.5. *Let $d \geq 2$, $k > d$ and set $\gamma := \frac{kd}{(k-d)(k-2)}$. Assume that $\mathbb{E}\|\xi_1\|^\gamma$ is finite and that $\gamma \geq 1$. If $Y_n := 2(\mathbb{E}[\int_{\mathbb{R}^d} |x|^\gamma \tilde{p}(dx)|\xi(n)])^{1/\gamma}$, then, for all $n < N$ and for P–almost every $\xi(n)$, one gets $Y := \sup_n Y_n < +\infty$ and*

$$W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) \leq \frac{1}{(N-n)^{1/k}}[\frac{4}{3} + 4 \cdot 3^{2k} \cdot 2^{d/2}(1 + Y_n)^{1/2}] + \frac{2n}{N}$$

Moreover,

$$P\left\{ W_1(\mathcal{L}_{\tilde{p}|\xi(n)}, \mathcal{L}_{\tilde{e}_N|\xi(n)}) > \epsilon \right\} \leq \frac{1}{\epsilon} \left[\frac{K}{(N-n)^{1/k}} + \frac{2n}{N} \right]$$

holds true for all $n < N$ with

$$K = \frac{4}{3} + 4 \cdot 3^{2k} \cdot 2^{d/2} \big(1 + 2(\mathbb{E}|\xi_1|^\gamma)^{1/\gamma}\big)^{1/2}.$$

Proof. Using (9) and applying the Jensen inequality two times, we obtain

$$\mathbb{E}[\Psi_k(\tilde{p})|\xi(n)] \leq \{2^d + 2^{d+1}(\mathbb{E}[\int_{\mathbb{R}^d} \|x\|^\gamma \tilde{p}(dx)|\xi(n)])^{1/\gamma}\}^{1/2}.$$

Combining Lemma 4 with this last inequality, Doob's martingale convergence theorem, Markov inequality and Jensen inequality concludes the proof. ♢

4.5. **Examples.** The application of the theorems of this section essentially require conditions on the moments of ξ_i. In the most common cases, the marginal distribution of each observation is available. Indeed, from a Bayesian point of view, the marginal distribution of each observation is usually treated as a prior guess of the mean of the unknown $\tilde{p}$. In the next three examples we review a few classical Bayesian nonparametric priors from this perspective.

Example 5 (Normalized random measures with independent increments)**.** Probably the most celebrated example of nonparametric priors is the Dirichlet process, see, for example, [10, 11]. A class of nonparametric priors which includes and generalizes the Dirichlet process is the class of the so called *normalized random measures with independent increments*, introduced in [31] and studied, e.g., in [26, 23, 24, 17, 32]. To define a normalized random measure with independent increments it is worth recalling that a random measure $\tilde{\mu}$ with independent increments on $\mathbb{R}^d$ is a random measure such that, for any measurable collection $\{A_1, \dots, A_k\}$ $(k \geq 1)$ of pairwise disjoint measurable subsets of $\mathbb{R}^d$, the random variable $\tilde{\mu}(A_1), \dots, \tilde{\mu}(A_k)$ are stochastically independent. Random measures with independent increments are completely characterized by a measure ν on $\mathbb{R}^d \times \mathbb{R}^+$ via their Laplace functional. More precisely, for every A in $\mathcal{B}(\mathbb{R}^d)$ and every positive λ one has

$$\mathbb{E}(e^{-\lambda\tilde{\mu}(A)}) = \exp\left\{-\int_{A\times\mathbb{R}^+}(1-e^{-\lambda v})\nu(dxdv)\right\}.$$

A systematic account of these random measures is given, for example, in [19]. Following [31], if $\int_{\mathbb{R}^d\times\mathbb{R}^+}(1-e^{-\lambda v})\nu(dxdv) < +\infty$ for every positive λ and $\nu(\mathbb{R}^d\times\mathbb{R}^+) = +\infty$, then one defines a *normalized random measure with independent increments* putting $\tilde{p}(\cdot) := \tilde{\mu}(\cdot)/\tilde{\mu}(\mathbb{R}^d)$. In point of fact, under the previous assumptions, $P\{\tilde{\mu}(\mathbb{R}^d) = 0\} = 0$; see [31]. The classical example is the Dirichlet process, obtained with $\nu(dxdv) = \alpha(dx)\rho(dv) = \alpha(dx)v^{-1}e^{-v}dv$, α being a finite measure on $\mathbb{R}^d$. Consider now a sequence $(\xi_i)_{i\geq 1}$ of exchangeable random variables driven by $\tilde{p}$. When $\nu(dxdv) = \alpha(dx)\rho(dv)$, then $P\{\xi_i \in A\} = \alpha(A)/\alpha(\mathbb{R}^d)$ for every $i \geq 1$. More generally,

$$P\{\xi_i \in A\} = \int_{\mathbb{R}^+}\phi(\lambda)\int_{A\times\mathbb{R}^+}e^{-\lambda u}u\nu(dxdu)d\lambda,$$

where

$$\phi(\lambda) := \exp\{-\int_{\mathbb{R}^k\times\mathbb{R}^+}(1-e^{-\lambda v})\nu(dydv)\}$$

see, e.g., Corollary 5.1 in [32]. Hence, $\mathbb{E}\|\xi\|^m < +\infty$ if and only if

$$\int_{\mathbb{R}^+}\phi(\lambda)\int_{\mathbb{R}^k\times\mathbb{R}^+}e^{-\lambda u}\|x\|^m u\nu(dxdu)d\lambda < +\infty.$$

Example 6 (Species sampling sequences and stick-breaking priors)**.** An exchangeable sequence of random variables $(\xi_n)_n$ is called a species sampling sequence (see [28]) if, for each $n \geq 1$,

$$P\{\xi_{n+1} \in A|\xi(n)\} = l_{0,n}\alpha(A) + \sum_{j=1}^{k(n)} l_{j,n}\delta_{\xi_j^*}(A) \qquad (A \in \mathcal{X})$$

and

$$P\{\xi_1 \in A\} = \alpha(A)$$

with the proviso that $\xi_1^*, \dots \xi_{k(n)}^*$ are the $k(n)$ distinct values of $\xi_1, \dots, \xi_n$ in the same order as they appear, $l_{j,n}$ $(j = 0, \dots, k(n))$ are non-negative measurable functions of $(\xi_1, \dots, \xi_n)$, and α is some non–atomic probability measure on $(X, \mathcal{X})$. See, among others, [14, 13, 27, 29]. Of

course, in this case, it a simple task to check conditions on the marginal distribution of each observation, since it coincides with α. A particular kind of random probability laws connected with the species sampling sequences are the so-called stick-breaking priors. Such priors are almost surely discrete random probability measures that can be represented as

$$\tilde{p}(\cdot) = \sum_{k=1}^{N} p_k \delta_{Z_k}(\cdot)$$

where $(p_k)_{k\geq 1}$ and $(Z_k)_{k\geq 1}$ are independent, $0 \leq p_k \leq 1$ and $\sum_{k=1}^{N} p_k = 1$ almost surely, and $(Z_k)_{k\geq 1}$ are independent and identically distributed random variable taking values in X with common probability α. Stick-breaking priors can be constructed using either a finite or infinite numbers of terms, $1 \leq N \leq +\infty$. Usually,

$$p_1 = V_1 \qquad p_k = (1 - V_1)(1 - V_2) \dots (1 - V_{k-1})V_k \qquad k \geq 2$$

where V_k are independent $Beta(a_k, b_k)$ random variables for $a_k > 0, b_k > 0$. See [16, 15]. It is clear that in this case

$$P\{\xi_i \in A\} = \alpha(A).$$

Example 7 (Pólya tree)**.** Let $X = \mathbb{R}$ and let E_j be the set of all sequences of 0s and 1s of length j. Moreover, set $E^* = \cup_j E_j$. For each n, let $\mathcal{T}_n = \{B_{\bar{\epsilon}} : \bar{\epsilon} \in E_n\}$ be a partition of $\mathbb{R}$ such that for all $\bar{\epsilon}$ in E^*, $B_{\bar{\epsilon}0}$, $B_{\bar{\epsilon}1}$ is a partition of $B_{\bar{\epsilon}}$. Finally let $\aleph = \{\alpha_{\bar{\epsilon}} : \bar{\epsilon} \in E^*\}$ be a set of nonnegative real numbers. A random probability $\tilde{p}$ on $\mathbb{R}$ is said to be a Pólya tree with respect to the partition $\mathcal{T} = \{\mathcal{T}_n\}_n$ with parameter $\aleph$ if

- $\{\tilde{p}\,(\,B_{\bar{\epsilon}0}|B_{\bar{\epsilon}0}) : \bar{\epsilon} \in E^*\}$ are a set of independent random variables
- for all $\bar{\epsilon}$ in E^* $\tilde{p}\,(B_{\bar{\epsilon}0}|B_{\bar{\epsilon}0})$ is $Beta(\alpha_{\bar{\epsilon}0}, \alpha_{\bar{\epsilon}1})$.

See [25, 20, 21]. Under suitable condition on $\aleph$, such a random probability does exist. See Theorem 3.3.2 in [12]. Moreover, if $(\xi_n)_{n\geq 1}$ is an exchangeable sequence with driving measure $\tilde{p}$, for any $B_{\bar{\epsilon}}$ with $\bar{\epsilon} = \epsilon_1\epsilon_2 \dots \epsilon_k$,

$$P\{\xi_n \in B_{\bar{\epsilon}}\} = \prod_{i=1}^{k} \frac{\alpha_{\epsilon_1\epsilon_2\dots\epsilon_i}}{\alpha_{\epsilon_1\epsilon_2\dots\epsilon_i 0} + \alpha_{\epsilon_1\epsilon_2\dots\epsilon_i 1}}.$$

See, e.g., Theorem 3.3.3 in [12]. In this case it is a difficult task to give explicit conditions for the existence of the moments of ξ_i. Nevertheless, Lavine suggests that, if the partitions has the form $F^{-1}(\sum \epsilon_i/2^i, \sum \epsilon_i/2^i + 1/2^i)$, F being a continuous distribution function, and

$$\frac{\alpha_{\epsilon_1\epsilon_2\dots\epsilon_i}}{\alpha_{\epsilon_1\epsilon_2\dots\epsilon_i 0} + \alpha_{\epsilon_1\epsilon_2\dots\epsilon_i 1}} = \frac{1}{2}$$

then $P\{\xi_n \leq x\} = F(x)$.

Acknowledgments

I am grateful to Eugenio Regazzini for providing much of the inspiration behind this paper. Moreover I want also to thank Luca Monno, who is a virtual coauthor of this paper, and Laura Sangalli for helpful comments. This work was partially supported by the IMATI (CNR - Pavia, Italy).

References

[1] K. S. Alexander. Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.*, **12**, 1041–1067, 1984.

[2] F. Bassetti and P.G. Bissiri. Finitary bayesian statistical inference through partitions tree distributions. *Sankhya*, to apper , 2006.

[3] F. Bassetti and P.G. Bissiri. Random partition model and finitary bayesian statistical inference. *Pubblicazioni IMATI-CNR*, (X-PV), 2006.

[4] F. Bassetti and E. Regazzini. The unsung de finetti's first paper about exchangeability. *Rendiconti di Matematica*, **28**, 2008.

[5] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. In *Memorie della Reale Accademia dei Lincei*, volume V of *IV*, 86–133, 1930.

[6] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Ann. Instit. H. Poincaré*, **7**, 1937.

[7] P. Diaconis and D. Freedman. Finite exchangeable sequences. *Ann. Probab.* **8** 745–764, 1980.

[8] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Statist*, **40**, 40–50, 1968.

[9] R. M. Dudley. *Real Analysis and Probability*, Cambridge University Press, Cambridge, 2002.

[10] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230, 1973.

[11] T. S. Ferguson. Prior distributions on spaces of probability measures. *Ann. Statist.*, **2**, 615–629, 1974.

[12] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer-Verlag, New York, 2003.

[13] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325 83–102, 244–245, 2005.

[14] B. Hansen and J. Pitman. Prediction rules for exchangeable sequences related to species sampling. *Statist. Probab. Lett.*, **46**, 251–256, 2000.

[15] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.*, **96**, 161–173, 2001.

[16] H. Ishwaran and L. F. James. Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankhyā*, **65**, 577–592, 2003.

[17] L. F. James. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Ann. Statist.*, **33**, 1771–1799, 2005.

[18] V. V. Kalashnikov and S. T. Rachev. *Mathematical Methods for Construction of Queueing Models*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[19] J. F. C. Kingman. Completely random measures. *Pacific J. Math.*, **21** 59–78, 1967.

[20] M. Lavine. Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, **20**, 1222–1235, 1992.

[21] M. Lavine. More aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, **22**, 1161–1176, 1994.

[22] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, Berlin, 1991.

[23] A. Lijoi, R. H. Mena, and I. Prünster. Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Stat. Inference Stoch. Process.*, **8**, 283–309, 2005.

[24] A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.*, **100**, 1278–1291, 2005.
[25] R. D. Mauldin and S. C. Williams. Reinforced random walks and random distributions. *Proc. Amer. Math. Soc.*, **110**, 251–258, 1990.
[26] L. E. Nieto-Barajas, I. Prünster, and S. G. Walker. Normalized random measures driven by increasing additive processes. *Ann. Statist.*, **32**, 2343–2360, 2004.
[27] J. Pitman. Exchangeable and partially exchangeable random partitions. *Proc. Roy. Soc. A.*, **102**, 1995.
[28] J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, probability and game theory*, volume 30 of *IMS Lecture Notes Monogr. Ser.*, 245–267. Inst. Math. Statist., Hayward, CA, 1996.
[29] J. Pitman. Poisson-Kingman partitions. In *Statistics and science: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes Monogr. Ser.*, pages 1–34. Inst. Math. Statist., Beachwood, OH, 2003.
[30] S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models.* John Wiley & Sons Ltd., Chichester, 1991.
[31] E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, **31**, 560–585, 2003.
[32] L. Sangalli. Some developments of the normalized random measures with independent increments. *Sankhyā*, **68**, 461–487, 2006.
[33] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, 1996.

Università degli Studi di Pavia, Dipartimento di Matematica, via Ferrata 1, 27100 Pavia, Italy

E-mail address: federico.bassetti@unipv.it